DOCUMENT RESUME

ED 231 838 TM 830 253

AUTHOR Rothman, M. L.; And Others
TITLE Generalizability Theory in Program Evaluation.
PUB DATE Aug 82
NOTE 12p.; Paper presented at the Annual Meeting of the American Psychological Association, (Washington, DC, August 23-27, 1982). Table 1 contains small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; Data Collection; Dental Schools; Educational Research; Error of Measurement; Higher Education; Interrater Reliability; *Learning Modules; *Program Evaluation; *Reliability; *Research Design; *Research Methodology; Statistical Analysis; Test Items
IDENTIFIERS Biomedical Computer Programs; Dental Auxiliary Education Project; *Generalizability Theory; Internal Consistency; Magnitude Estimates; Repeated Measures Design; *Variance (Statistical)

ABSTRACT

A practical application of generalizability theory, demonstrating how the variance components contribute to understanding and interpreting the data collected to evaluate a program, is described. The evaluation concerned 120 learning modules developed for the Dental Auxiliary Education Project. The goals of the project were to design, implement, and evaluate an innovative curriculum for dental assistants and dental hygienists. A repeated measures design was developed in which students were nested within schools which were crossed with modules and items. Generalizability theory allows for direct comparison of the relative magnitude of the sources of variance in the study design, as well as calculation of a variety of useful generalizability coefficients indicating the relative degree of accuracy with which the investigator can generalize across factors. It allows an assessment of the efficiency of the design in that it is possible to assess changes in generalizability coefficients with changes in numbers of subjects, items, groups, etc. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED231838

GENERALIZABILITY THEORY IN PROGRAM EVALUATION

by

M.L. Rothman, J.C. Wintz, P.D. Peckham, C.W. Dohner

Division of Research in Medical Education
University of Washington
Seattle, Washington

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M.L. Rothman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 253

Paper presented at the Annual Meeting of the American Psychological Association, Washington, D.C., 1982.

Traditionally reliability has been defined as the extent to which which the results of a measurement were due to systematic sources of variance. More recently the term has been used to indicate the adequacy with which one can generalize from one observation to a universe of observations. (van de Kamp, 1976)

This issue becomes especially relevant in program evaluation where there are frequently multiple sources of variance. For example, students or clients are frequently nested within classes or groups and thus it is important for the evaluator to be able to assess the impact of class or group membership on the rating of the program. One method of assessing this impact is through generalizability theory, an approach developed by Cronbach, Glesser, Nanda and Rajaratnam (1972), and more recently elaborated and applied by Gilmore, Kane and Naccarato (1976, 1978), Brennan (1977), Kane and Brennan (1978), Rentz (1980). This approach allows the investigator to estimate various correlation coefficients which provide an index of the accuracy of generalizing across levels of a factor, including estimates of internal consistency and interrater reliability. Also useful, is the generation of the various components for each source of variance in the design. Examination of these components allows the evaluator to assess the relative contribution of each component to the total variance which, in turn, provides information about adjusting the design for increased efficiency.

Frequently articles dealing with generalizability theory are quite technical and spend much time addressing the issue of calculation of variance components. The technical jargon, plus the tediousness of calculating the variance components has led to infrequent utilization of generalizability theory by the practically-oriented investigator. However, it is possible to calculate the components of variance utilizing the Biomedical Computer Programs, BMDP8V (1979) with a balanced design.

Once these components are calculated, the calculation of generalizability coefficients is relatively straightforward. A practical application of generalizability theory, demonstrating how the variance components contribute to understanding and interpreting the data collected to evaluate a program is described below.

The Dental Auxiliary Education Project is a Kellogg Foundation funded project. The goals of the project were to design, implement and evaluate an innovative curriculum for dental assistants and dental hygienists. As the evaluators of the project, one of our goals was to evaluate 120 learning modules developed by the project staff. The original evaluation plan called for each student in each school to evaluate each module ($n_{schools} = 7$, $n_{students/school} = 10$ to 40). This information was used to aid the module writers in revising the modules and as a final product evaluation for those modules that were not revised. The question arose as to the similarity of student ratings of modules across schools (i.e. was it sufficient to present the module writer with a single item mean or should responses be analyzed by school). Due to the limitation of time and money, a second question arose, that is, was it possible to make a statement about the modules as a set or would it be necessary to consider each of the modules separately in making an evaluative statement (i.e. was the quality of modules consistent across modules)?

A generalizability study allowed us to address these questions. The design developed to respond to the above questions was a repeated measures design in which students were nested within schools which were crossed with modules and items. This design is depicted in Figure 1.

Insert Figure 1

An issue of primary importance to any evaluation is the reliability of the instrument used to obtain the data. Estimates of reliability usually include estimates of internal consistency of the items and interrater reliability. It is possible to estimate these coefficients for the design depicted in Figure 1, utilizing the variance components

shown in Table 1. Calculation of the two generalizability coefficients is shown below:

$$\rho^2_{(i)} = \frac{\sigma^2_m + \frac{1}{n_a}\sigma^2_{am} + \frac{1}{n_{sa}}\sigma^2_{ms(a)}}{\sigma^2_m + \frac{1}{n_a}\sigma^2_{am} + \frac{1}{n_i}\sigma^2_{mi} + \frac{1}{n_{sa}}\sigma^2_{ms(a)} + \frac{1}{n_{ai}}\sigma^2_{ami} + \frac{1}{n_{sia}}\sigma^2_{smi(a)}} \quad (1)$$

$$\rho^2_{(s:a)} = \frac{\sigma^2_m + \frac{1}{n_i}\sigma^2_{mi}}{\sigma^2_m + \frac{1}{n_a}\sigma^2_{am} + \frac{1}{n_i}\sigma^2_{mi} + \frac{1}{n_{sa}}\sigma^2_{sm(a)} + \frac{1}{n_{ai}}\sigma^2_{ami} + \frac{1}{n_{sia}}\sigma^2_{smi(a)}} \quad (2)$$

Equation 1 represents an estimate of the internal consistency of the items, i.e. an estimate of the correlation between this set of items and another random set of items rated by the same students. In this example the letter in parentheses indicates the factor we want to generalize across while modules is the object of measurement or the factor we want to make decisions about and students within schools is a fixed factor. The generalizability coefficient for this equation with n_a=3, $n_{s:a}$=10 and n_i=11 and n_m=5 is .89, indicating a high degree of internal consistency for items.

Equation 2 represents an estimate of interrater reliability, i.e. an estimate of the correlation between the ratings of the modules between this random sample of students and another random sample of students using the same items. Modules is again the object of measurement, while students within schools is considered a random factor and items are fixed. The generalizability coefficient is .12. Obviously the size of this coefficient does not allow generalization across raters with any degree

of precision. Estimates of the impact of increasing either the number of students within schools or the number of schools (see Figure 2) indicates that this is not a useful strategy for increasing estimates of interrater reliability. For instance, with 3 schools and 20 students per school, the generalizability coefficient is .15 (assuming the original sample was representative). Even with 7 schools and 30 students per school the generalizability coefficient is only .26. An examination of the variance components shows a relatively large school by module interaction, indicating that the ratings of the modules were not consistent across schools. Separation of the schools into types (dental assisting and dental hygiene) yields a generalizability coefficient of .62 with 10 students per school and .71 with 15 students per school. It becomes clear that the ratings of modules cannot be generalized across types of schools, although ratings within types of schools can achieve an acceptable level of interrater reliability, at least for the dental hygiene schools.

Insert Figure 2

The question of homogeneity of the modules with respect to quality was also addressed through examination of the variance components. It can be seen that the relative magnitude of the component due to modules is quite small, indicating that differences in student perceptions of the quality of the modules summed over items and schools was quite similar. This is demonstrated by the size of the generalizability coefficient for student ratings of this set of modules and another set of modules using the same items (ρ^2 = .82 for all 3 schools and .89 for the dental hygiene schools only). These coefficients were calculated using equation 3.

$$\rho^2_{(m)} = \frac{\sigma^2_{s:a} + \frac{1}{n_i}\sigma^2_{si(a)}}{\sigma^2_{s:a} + \frac{1}{n_m}\sigma^2_{sm(a)} + \frac{1}{n_i}\sigma^2_{si(a)} + \frac{1}{n_{mi}}\sigma^2_{smi(a)}} \qquad (3)$$

In summary, it can be seen from the preceeding study that generalizability theory can be a very useful tool for the program evaluator. It allows for direct comparison of the relative magnitude of the sources of variance in the study design, as well as, calculation of a variety of useful generalizability coefficients indicating the relative degree of accuracy with which the investigator can generalize across factors. In addition, it allows an assessment of the efficiency of the design in that it is possible to assess changes in generalizability coefficients with changes in numbers of subjects, items, groups, etc.

REFERENCES

Brennan, R.L. Generalizability analysis: Principles and procedures. ACT Technical Bulletin, No. 26. Iowa City, Iowa: The American College Testing Program, 1977.

Cronbach, L.J., Glesser, G.C., Nanda, H. and Rajartnam, N. The dependability of behavioral measurements; Theory of generalizability for scores and profiles. New York: Wiley, 1972.

van der Kamp, L.J.T. Generalizability and educational measurement. Advances in psychological and educational measurement, (Eds.) D.M.N. de Gruijter and L.J.T. van der Kamp. New York: Wiley, 1976.

Gilmore, G.M., Kane, M.T. and Naccarato, R.W. The generalizability of student instructional ratings: General theory and application to the University of Washington instructional assessment system. Educational Assessment Center Project:506. University of Washington, Seattle, Washington, 1976.

Kane, M.T. and Brennan, R.L. The generalizability of class means. Review of Educational Research, 1977, 47, 267-292.

Rentz, R.R. Rules of thumb for estimating reliability coefficients using generalizability theory. Educational and Psychological Measurement, 1980, 40, 575-593.

TABLE 1

SUMMARY OF RANDOM EFFECTS ANOVA AND ESTIMATES OF VARIANCE COMPONENTS

Source	Notation	Degrees of Freedom	n	Expected Mean Square	Estimates of Variance Components *	Estimates of Variance Components +
School	a	(a-1)	3	$\sigma^2_{smi(a)} + n_s\sigma^2_{aim} + n_m\sigma^2_{si(a)} + n_i\sigma^2_{sm(a)} + n_{sm}\sigma^2_{ai} + n_{is}\sigma^2_{am} + n_{mi}\sigma^2_{s(a)} + n_{mis}\sigma^2_{a}$	.07205	-.01421 #
Module	m	(m-1)	5	$\sigma^2_{smi(a)} + n_s\sigma^2_{ami} + n_i\sigma^2_{sm(a)} + n_{s(a)}\sigma^2_{mi} + n_{si}\sigma^2_{am} + n_{si(a)}\sigma^2_{m}$	.00397	.01397
Item	i	(i-1)	11	$\sigma^2_{smi(a)} + n_s\sigma^2_{ami} + n_m\sigma^2_{si(a)} + n_{s(a)}\sigma^2_{mi} + n_{sm}\sigma^2_{ai} + n_{sm(a)}\sigma^2_{i}$	.09151	.14682
Student within school	s:a	a(s-1)	10	$\sigma^2_{smi(a)} + n_m\sigma^2_{si(a)} + n_i\sigma^2_{sm(a)} + n_{mi}\sigma^2_{s(a)}$	.12545	.15998
School x module	am	(a-1)(m-1)		$\sigma^2_{smi(a)} + n_s\sigma^2_{ami} + n_i\sigma^2_{sm(a)} + \sigma^2_{si}\sigma^2_{am}$	.04882	-.00576 #
School x item	ai	(a-1)(i-1)		$\sigma^2_{smi(a)} + n_s\sigma^2_{ami} + n_m\sigma^2_{si(a)} + n_{sm}\sigma^2_{ai}$	.02268	.01447
Module x item	mi	(m-1)(i-1)		$\sigma^2_{smi(a)} + n_s\sigma^2_{ami} + n_{s(a)}\sigma^2_{mi}$	.00177	.01412
Student x module within school	sm:a	a(s-1)(m-1)		$\sigma^2_{smi(a)} + n_i\sigma^2_{sm(a)}$	.10469	.06311
Student x item within school	si:a	a(s-1)(i-1)		$\sigma^2_{smi(a)} + n_m\sigma^2_{si(a)}$	.07179	.08057
School x module x item	ami	(a-1)(m-1)(i-1)		$\sigma^2_{smi(a)} + n_s\sigma^2_{ami}$	.00938	-.00441 #
Student x module x item within school	smi:a	a(s-1)(m-1)(i-1)		$\sigma^2_{smi(a)}$	.36304	.32987

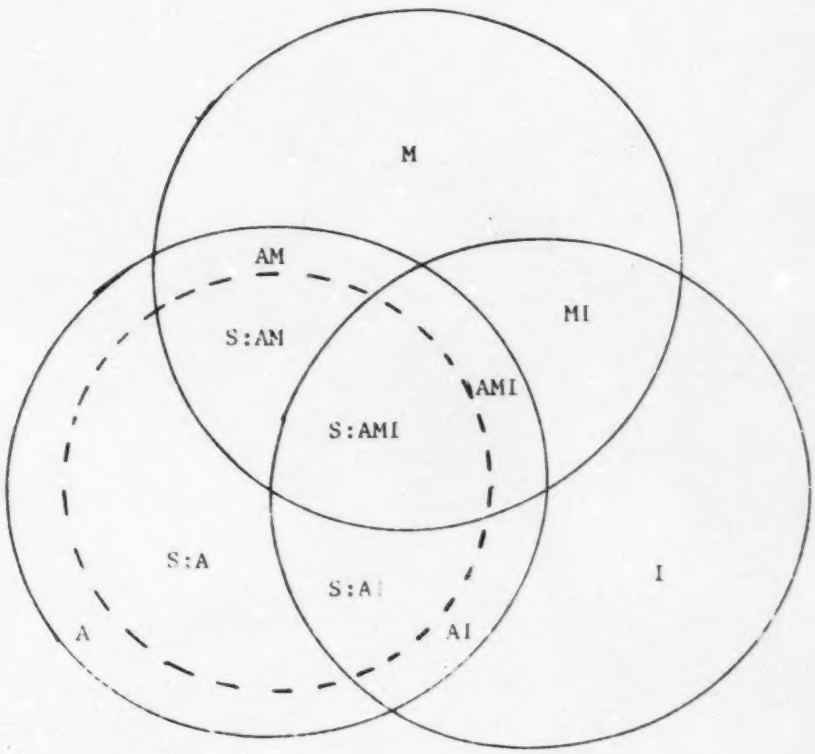
* estimates of variance components with n=3 schools (summed across types of schools)

\+ estimates of variance components with n=2 schools (one type of school)

\# negative variance components are treated as zeros following the suggestion of Cronbach, et al. (1972)

Figure 1

Venn Diagram of ANOVA Design



A = schools
M = modules
I = items
S:A = students within schools
AM = school x module
AI = school x item
MI = module x item
S:AM = students within school by module
S:AI = students within school by item
AMI = school x module x item
S:AMI = students within school x module x item

Figure 2

Estimates of interrater reliability for various numbers of schools and students per school